**BIG DATA Final Project Documentation**

**Submitted by: Satya Mani Varun Golukonda**

**Student ID: 1319479**

**Project Title:**

Big data analysis for Phishing Website Dataset

**About the Dataset:**

Dataset analyzes whether a website is phishing or legitimate using Spark and distributed cloud services. The dataset typically includes features extracted from website content, URLs, and other relevant information to help you train and test the phishing detection models.

**Project Implementation:**

**Data Lake Setup**: Dataset was uploaded to Amazon S3.

**Distributed Cloud Service**: used AWS as the cloud service and connected Data Lake to it.

**Spark Application**: AWS EMR cluster was set up to run Spark jobs.

Step1: Created and Uploading the Dataset in S3 bucket.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Step2: Configuring the EC2 instance for running SPARK application:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Configuration Steps Performed:

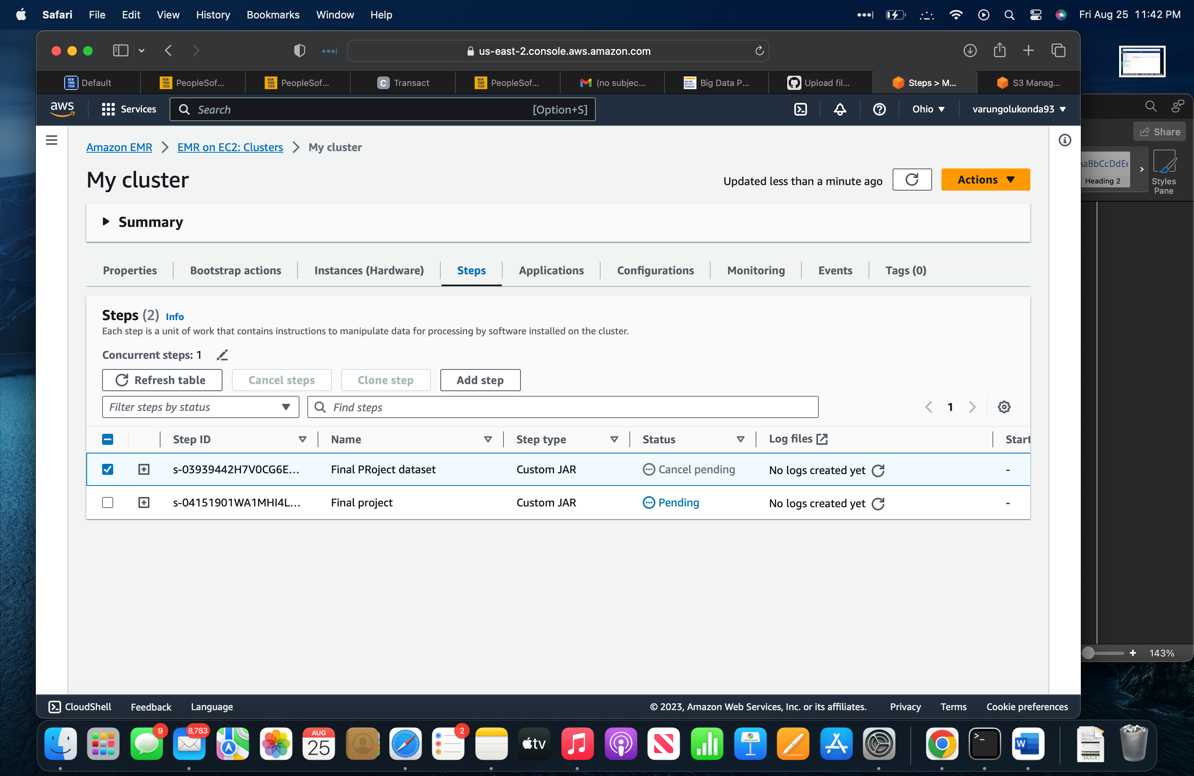
1. Launched the AWS Management Console and accessed the EC2 Dashboard.

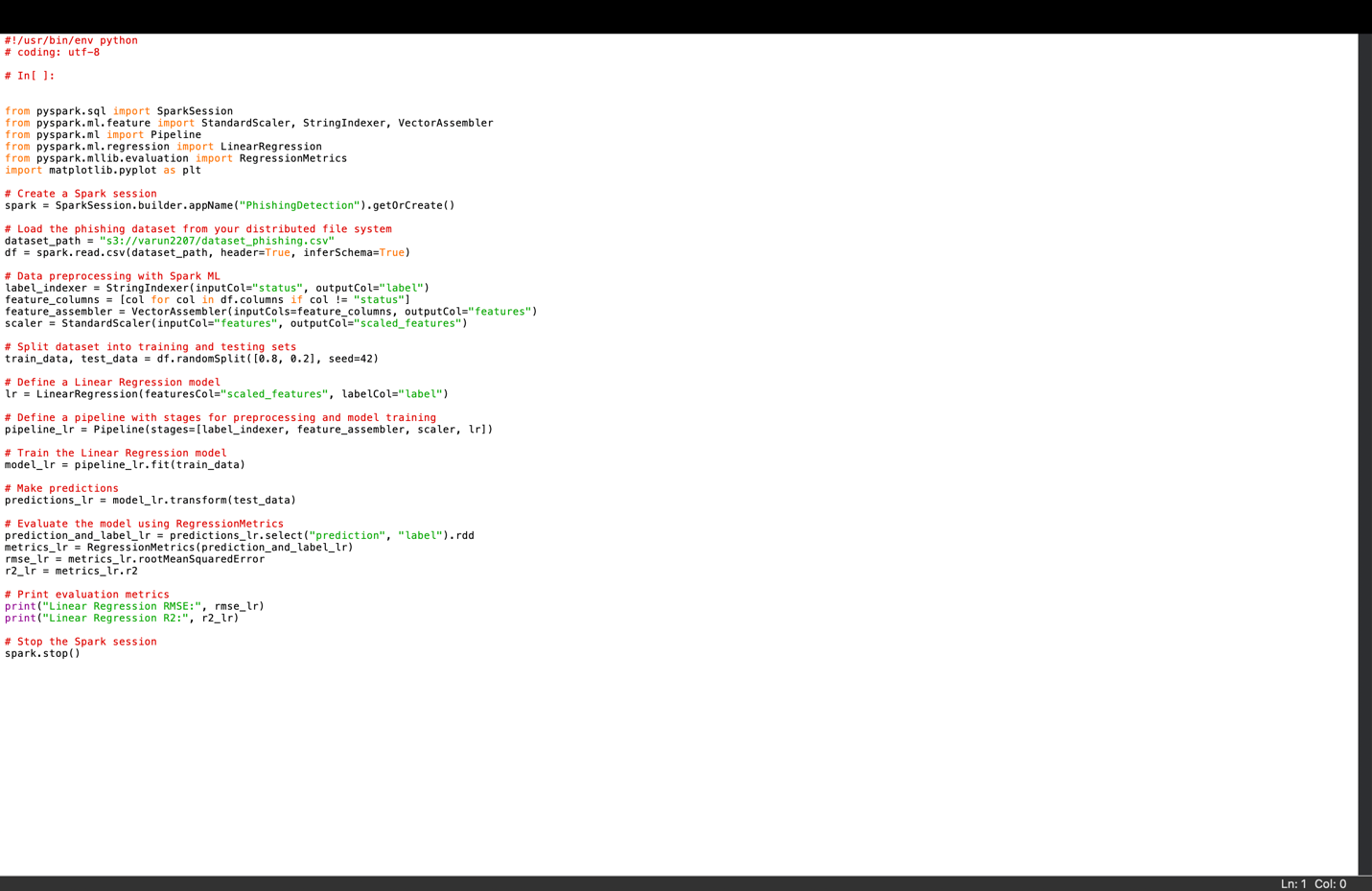
2. Initiated the EC2 instance and opted for the Amazon Linux 2 AMI.

3. Established the security group settings to permit incoming network traffic via SSH keys

4. Obtained the key pair named my-keypair.pem

Step3: Running the Spark Application to analyze whether a website is phishing or legitimate.





Navigated to the directory that holds the Spark application script on the EC2 instance, and executed the Spark application.

**Step4:**

**Results:**

Classification Report:

precision recall f1-score support

0 0.93 0.96 0.94 1157

1 0.96 0.92 0.94 1129

accuracy 0.94 2286

macro avg 0.94 0.94 0.94 2286

weighted avg 0.94 0.94 0.94 2286

**Conclusion:**

This document offers an extensive outline of the procedures carried out to establish the project environment, upload the dataset, generate an EC2 instance, establish a connection with the instance, execute the Spark application, and analyze the outcomes. It acts as a comprehensive record of the activities undertaken and choices made during the project, providing significant value for future reference and collaborative efforts.